DOCUMENT RESUME

ED 361 348 TM 020 399

AUTHOR Bush, M. Joan; Schumacker, Randall E.
TITLE Quick Norms with Rasch Measurement.
PUB DATE Apr 93
NOTE 24p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Computer Simulation; Item Response Theory; Norm Referenced Tests; *Sample Size; Sampling; *Statistical Distributions; Test Interpretation; Test Length; *Test Norms
IDENTIFIERS *Mean (Statistics); *Rasch Model

ABSTRACT

The feasibility of quick norms derived by the procedure described by B. D. Wright and M. H. Stone (1979) was investigated. Norming differences between traditionally calculated means and Rasch "quick" means were examined for simulated data sets of varying sample size, test length, and type of distribution. A 5 by 5 by 2 design with a total of 50 experiments was used, and each experiment was replicated 100 times. The BIGSTEPS Rasch calibration program was used to analyze each of the 5,000 data sets. Quick norms were calculated using programs from the Statistical Package for the Social Sciences. The Rasch quic norms procedure yielded means that were equivalent to traditionally calculated means for tests with a minimum of 30 items given to groups of 50 examinees or more, for both normally and uniformly distributed item difficulties. The methods were not equivalent with tests with 10 items. Sample size was not a factor in determining differences between the two methods. The quick norm procedure is recommended in cases where there is an existing bank of Rasch calibrated items. Its simplicity and ease of use makes it advantageous. Three tables present simulation data. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 361 348

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

JOAN BUSH

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Quick Norms with Rasch Measurement

M. Joan Bush
Professional Development Institute

&

Randall E. Schumacker
University of North Texas

Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, Georgia, April 12-16, 1993.

TM020399

BEST COPY AVAILABLE

Introduction

The classical test model, used by measurement specialists for many years, is simple to understand and makes use of several well-known statistics; however, there are several problems or shortcomings of educational and psychological tests based upon classical true score theory (Allen & Yen, 1979; Crocker & Algina, 1986). With a norm-referenced test based upon classical true score theory, a person's score would, in most cases, be different if he or she took a test with a different group of people or with a different set of items. Wright and Stone (1979) said it best:

> If all of a specified set of items have been tried by a child you wish to measure, then you can obtain his percentile position among whatever groups of children were used to standardize the test. But how do you interpret this measure beyond the confines of that set of items and those groups of children? Change the children and you have a new yardstick. Change the items and you have a new yardstick again. Each collection of items measures an ability of its own. Each measure depends for its meaning on its own family of test takers. How can we make objective mental measurements and build a science of mental development when we work with rubber yardsticks? (p. xi)

Thus, to interpret and make sense of a person's score based upon classical true score theory, we must know the group and the specific test that the examinee took.

Rasch measurement has several advantages over traditional norm-referenced testing. First, Rasch measurement results in sample-free item calibrations and test-free person measurements (Rasch, 1966; Wright, 1967). Second, Rasch measurement estimates are more precise than traditional estimates since standard errors

of measurement are calculated for each examinee and each test. As for classical test theory, "the same standard error of measurement usually is used for all true scores" (Allen & Yen, 1979, 253). In addition, Rasch parameter estimators are unbiased, consistent, efficient, and sufficient (Andrich, 1988; Wright & Stone, 1979).

Norm-Referencing with Rasch Measurement

With traditional norm-referencing, normative scores provide information about the examinee's test performance as compared to the distribution of scores from a reference group or norm sample. Thus, for scores to be meaningful, "a particular set of norms needs to be relevant to the desired interpretation of examinee performance" (Peterson, Kolen, Hoover, 1989, p. 236). Examples of norms include national norms, national subgroup norms, local norms, user norms, convenience norms, norms for school averages, item and skill norms. Norms are conveyed by summary statistics including means, standard deviations, and percentile ranks. Other common normative scores include linear z-scores, normalized z-scores, derived scores (T, IQ, and NCEs), stanines, scaled scores, and grade and age equivalents (Crocker & Algina, 1986; Peterson et al., 1989).

Norms can be developed based upon a bank of Rasch calibrated items. Wright and Stone (1979) described a procedure to estimate quick norms based upon the Rasch model. According to Wright and Stone (1979),

> norming a variable in the Rasch approach takes much less data than norming a test . . . once the variable is normed, then all possible scores from all possible tests drawn from the calibrated bank are automatically norm-referenced through the variable (p. 26).

To determine the mean and standard deviation for each cell (or sample) in the normative sampling plan, quick norms can be estimated from frequency data on the calibrated items without measuring each person individually. Although the mean and standard deviation can be estimated from a random sample of approximately 100 persons on two items, Wright and Stone recommend a longer test of ten to fifteen items. The procedure is as follows:

1. For each cell or sample in the norming study, a set of calibrated items is selected in which items are sufficiently spaced in difficulty to cover the expected dispersion of abilities of the particular sampling cell. The result is an individually tailored norming test for each sampling cell.

2. The selected set of items is administered to a random sample of persons from the specified sampling cell.

3. The frequency of persons who succeed on each item is calculated.

4. The natural log odds for correct answers is calculated for each item using the formula $h_i=\ln[(s_i/N-s_i)]$ for each of the K items where

s_i = number of persons succeeding on an item,

h_i = natural log odds of the correct answers, and

N = # of persons in the sampling cell.

5. The log odds correct answers are regressed on the

associated item difficulties over the K items to obtain the intercept (A) and slope (C) of the least squares straight line.

6. The estimated population mean and standard deviation of the sampling cell's abilities are calculated using the formulas

$$M = -A/C$$

$$SD = 1.7[(1-C^2)/C^2]^{1/2}$$

where

M = estimate of population mean,

SD = estimate of population standard deviation,

A = intercept, and

C = slope.

No studies have been documented related to the quick norm procedure described by Wright and Stone (1979). As a result, this study investigated the feasibility of quick norms based upon Rasch measurement. Norming differences between traditionally calculated means and Rasch "quick" means were examined for simulated data sets of varying sample size, test length, and type of distribution.

Methods and Procedures

Synthetic data sets containing raw score data for dichotomously scored items were used in the investigation. Data were simulated to vary in test length, sample size, and distribution. A 5 by 5 by 2 design with a total of 50 experiments was used. Five (5) test lengths (10, 20, 30, 40, and 50 items), 5 sample sizes (50, 100, 200, 300, and 400 persons), and 2 types of item difficulty distributions (normal and uniform) were completely crossed. Each experiment was replicated 100 times.

The experiments will be referred to by number as shown in Table 1. For example, Experiment 1 consisted of a 10-item test with normally distributed item difficulties given to 50 examinees. For each experiment, there were 100 datasets.

Construction of Simulated Data Sets

The data sets were constructed using SIMTEST 2.1 (Luppescu, 1992). SIMTEST generates dichotomous examinee responses based upon several values specified by the user. SIMTEST allows specification of population parameters for the item and person measures. The following parameters were used in this study.

1. The item difficulty distribution was set to either uniform or normal.

2. Item difficulty parameters for the uniformly distributed data sets had a range of 3 to 4 logits from easiest to most difficult. The mean was set to 0 and the standard deviation was set to 2. As a result, the item difficulties were uniformly distributed with 95% of the difficulties lying between -2 and +2.

3. Item difficulty parameters for the normally distributed data sets also had a range of 3 to 4 logits. The mean was set to 0 and the standard deviation was set to 1. As a result, the item difficulties were normally distributed with 95% of the item difficulties ranging from -2 to +2 standard deviations with most of them near 0.

4. The mean of person ability measures was set at 0 for both the normal and uniform item difficulty distributions. As a result, the person abilities were normally distributed for all of the

Table 1

Definition of Experiments

Experiment	Distribution	No. of Items	No. of Persons
1	Normal	10	50
2	Normal	20	50
3	Normal	30	50
4	Normal	40	50
5	Normal	50	50
6	Normal	10	100
7	Normal	20	100
8	Normal	30	100
9	Normal	40	100
10	Normal	50	100
11	Normal	10	200
12	Normal	20	200
13	Normal	30	200
14	Normal	40	200
15	Normal	50	200
16	Normal	10	300
17	Normal	20	300
18	Normal	30	300
19	Normal	40	300
20	Normal	50	300

Experiment	Distribution	No. of Items	No. of Persons
21	Normal	10	400
22	Normal	20	400
23	Normal	30	400
24	Normal	40	400
25	Normal	50	400
26	Uniform	10	50
27	Uniform	20	50
28	Uniform	30	50
29	Uniform	40	50
30	Uniform	50	50
31	Uniform	10	100
32	Uniform	20	100
33	Uniform	30	100
34	Uniform	40	100
35	Uniform	50	100
36	Uniform	10	200
37	Uniform	20	200
38	Uniform	30	200
39	Uniform	40	200
40	Uniform	50	200
41	Uniform	10	300
42	Uniform	20	300

Experiment	Distribution	No. of Items	No. of Persons
43	Uniform	30	300
44	Uniform	40	300
45	Uniform	50	300
46	Uniform	10	400
47	Uniform	20	400
48	Uniform	30	400
49	Uniform	40	400
50	Uniform	50	400

datasets.

Other values were set as follows:

1. The threshold for guessing was set at 0 for each of the data sets.

2. The slope of all items was set at 1 since this is an assumption of the Rasch model.

3. No bias measures were used in this study.

Rasch Analysis

The BIGSTEPS Rasch calibration program was used to analyze each of the 5000 data sets (Linacre, 1992). A BIGSTEPS program was written to output the item difficulty measures and the frequency of persons who answered each item correctly.

Calculation of Rasch Quick Norms

The item difficulties and frequencies of persons who answered each item correctly were used to calculate the quick norms as follows:

1. The frequency of persons who succeeded on each item was computed.

2. The natural log odds for correct answers was calculated for each item using the formula $h_i=\ln[(s_i/N-s_i)]$ where:

s_i = number of persons who succeeded on an item,

h_i = natural log odds of the correct answers, and

N = # of persons in the sampling cell.

3. The log odds correct answers were regressed on the associated item difficulties over the K items to obtain the intercept (A) and slope (C) of the least squares straight line.

4. In addition, the estimated population mean and standard deviation of the sampling cell's abilities were calculated using the following formulas:

$M = -A/C$, and

$SD = 1.7[(1-C^2)/C^2]^{1/2}$

where

M = estimate of population mean,

SD = estimate of population standard deviation,

A = intercept, and

C = slope.

An SPSSX program was used to perform steps 1 through 3 of the quick norm procedure. The data used in the SPSSX program came directly

from the BIGSTEPS output file which contained the item difficulty measures and the number of examinees who answered each item correctly. A second SPSSX program, using the slopes and intercepts from the previous SPSSX programs, was written to perform the calculations for steps 4 and 5 of the quick norm procedure.

Traditional Analysis

The traditional mean and standard deviation were computed for each of the 5000 datasets. An SPSSX program was written to perform these calculations.

Assessment of the Quick Norm Procedure

The overall population mean and standard error of the mean were calculated for the traditionally calculated means and for the Rasch quick means for each of the 50 experiments. The traditionally calculated means were rescaled to a mean of zero and a standard deviation of 1 to place the traditionally calculated means and the Rasch quick means on the same scale. An SPSSX program was written to compute the population mean and standard error of the mean for each experiment.

To find out if there were significant differences between the means calculated by the 2 methods, an independent t-test was computed. An SPSSX program was used to compute the t-tests for the 50 experiments.

Results

Table 2 contains the population means, standard errors of the means, and absolute mean differences between traditional and Rasch

quick norms for the tests that had normally distributed item difficulties. The first section of the table includes the information for the first 5 datasets which consisted of tests with normally distributed item difficulties given to 50 persons. The difference between the 2 types of means for the 10-item test was significant ($p<.0001$). The absolute differences between the traditionally calculated and Rasch quick means for the 20- and 30-item tests were closer than those of the 10-item test with approximate absolute differences of .011. As for the 40- and 50-item tests, the differences between the 2 types of means were hardly discernible with absolute differences of .0071 and 0026, respectively. The standard errors of the means were somewhat larger for the Rasch quick norms than for the traditionally calculated norms for all 5 test lengths.

Datasets 6 through 10 reflected simulated tests with normally distributed item difficulties given to 100 persons. The results in the second section of Table 2 show that the 2 means for the 10-item test were significantly different ($p<.0001$) with an absolute difference of .0723. The traditional mean and Rasch quick mean for the 20-item test were close with an absolute difference of .0163. As for the 30-, 40-, and 50-item tests, the differences between the traditional and Rasch quick means were very small, ranging from .0054 to .0007. In all cases, the standard errors of the means for the traditional means were slightly smaller than the standard errors of the means for the Rasch quick means.

Table 2

Comparison of Traditional and Rasch Quick Norms
Normal Distribution

No. of Items	Traditional Norms		Rasch Quick Norms		μ Difference
	μ	$\sigma_{\bar{x}}$	μ	$\sigma_{\bar{x}}$	
			50 Persons		
10	.0200	.0131	-.0881	.0166	.1081*
20	.0027	.0131	-.0080	.0148	.0107
30	-.0026	.0135	-.0132	.0157	.0106
40	-.0066	.0128	-.0137	.0149	.0071
50	-.0113	.0123	-.0139	.0139	.0026
			100 Persons		
10	-.0144	.0110	-.0867	.0130	.0723*
20	-.0029	.0102	-.0192	.0120	.0163
30	.0147	.0097	.0093	.0107	.0054
40	.0154	.0104	.0167	.0116	.0013
50	-.0015	.0095	-.0022	.0107	.0007
			200 Persons		
10	-.0027	.0066	-.0963	.0087	.0936*
20	-.0012	.0065	-.0190	.0072	.0178
30	.0044	.0064	.0000	.0074	.0044
40	.0042	.0075	.0037	.0081	.0005
50	.0091	.0065	.0096	.0073	.0005
			300 Persons		
10	-.0075	.0057	-.1002	.0063	.0927*
20	.0076	.0062	-.0081	.0071	.0157
30	-.0050	.0062	-.0110	.0067	.0060
40	-.0071	.0059	-.0097	.0065	.0026
50	.0116	.0057	-.0126	.0063	.0010
			400 Persons		
10	.0038	.0066	-.0886	.0071	.0924*
20	.0086	.0052	-.0073	.0061	.0159**
30	-.0021	.0053	-.0068	.0059	.0047
40	.0048	.0048	.0028	.0054	.0020
50	-.0075	.0041	-.0088	.0046	.0013

* $p<.0001$
** $p<.05$

The population means, standard errors of the means, and absolute mean differences for datasets 11 to 15 can be seen in section 3 of Table 2. These simulated datasets included the responses of 200 persons on tests with normally distributed item difficulties. Overall, the traditional means and Rasch quick means became increasingly closer as test length was increased. The difference between the 2 types of means for the 40- and 50-item tests was hardly noticeable with both having absolute differences of .0005. In contrast, the traditional mean and Rasch quick mean were significantly different ($p<.0001$) for the 10-item exam with an absolute difference of .0936. As before, the standard errors of the means were slightly larger for the Rasch quick norms than for the traditionally calculated norms.

Datasets 16 through 20 included simulated data for 300 persons who took tests with normally distributed item difficulties. As can be seen in section 4 of Table 2, the most pronounced difference between the 2 methods of calculating means was for the 10-item test. In this case, there was a significant difference ($p<.0001$) between the traditional and Rasch quick means. Once again, the traditional and Rasch quick means became progressively closer as the test length was increased with the 50-item test showing the smallest difference of .001. In all cases, the standard errors of the means for the traditional means were smaller than those of the Rasch quick means.

The last section of Table 2 shows the population means, standard errors of the means, and absolute differences between

the traditional and Rasch quick means for datasets 21 through 25. These simulated datasets reflect the responses of 400 examinees to tests containing normally distributed item difficulties. Like datasets 1 to 20, the traditionally calculated mean of .0038 and the Rasch quick mean of -.0886 for the 10-item exam were significantly different ($p<.0001$), resulting in an absolute difference of .0924. In addition, the difference between traditional and Rasch quick means was significant ($p<.05$) for the 20-item test. Once again, the traditional and Rasch quick means showed decreasing absolute differences as test length increased, ranging from .0147 for the 30-item test down to .0013 for the 50-item test. The Rasch standard errors of the means were slightly larger than the traditional standard errors of the means for all test lengths.

Table 3 contains the population means, standard errors of the means, and absolute mean differences between traditional and Rasch quick norms for the tests that had uniformly distributed item difficulties. The means and standard errors of the means for datasets 26 to 30 are shown in the first section of Table 3. These sets of data had a sample size of 50 and reflected tests with uniformly distributed item difficulties. There was a significant difference ($p<.01$) between the traditional and Rasch quick means for the 10-item test. The traditional and Rasch quick means for the 30- and 50-item tests were very close, yielding absolute differences of .0022 and .0024, respectively. The smallest difference between the 2 types of means was for the

Table 3

Comparison of Traditional and Rasch Quick Norms
Uniform Distribution

Test Length	Traditional Norms		Rasch Quick Norms		μ Difference
	μ	$\sigma_{\bar{x}}$	μ	$\sigma_{\bar{x}}$	
			50 Persons		
10	-.0010	.0152	-.0769	.0172	.0759**
20	.0067	.0143	-.0053	.0176	.0129
30	.0157	.0143	.0135	.0173	.0022
40	.0095	.0127	.0091	.0156	.0004
50	.0168	.0135	.0192	.0163	.0024
			100 Persons		
10	.0249	.0099	-.0689	.0117	.0938*
20	-.0149	.0089	-.0308	.0106	.0159
30	-.0079	.0093	-.0111	.0111	.0032
40	.0112	.0103	.0112	.0124	.0000
50	.0034	.0091	.0033	.0108	.0001
			200 Persons		
10	-.0018	.0066	-.0817	.0083	.0799*
20	.0091	.0063	-.0035	.0076	.0126
30	-.0060	.0062	-.0098	.0073	.0038
40	.0087	.0053	.0089	.0064	.0002
50	-.0057	.0067	-.0067	.0080	.0010
			300 Persons		
10	.0028	.0056	-.0717	.0076	.0745*
20	.0086	.0047	-.0021	.0058	.0107
30	.0077	.0055	.0048	.0067	.0029
40	-.0010	.0061	-.0031	.0073	.0021
50	.0130	.0059	.0144	.0070	.0014
			400 Persons		
10	.0082	.0056	-.0650	.0065	.0732*
20	.0007	.0049	-.0121	.0058	.0128
30	-.0023	.0048	-.0052	.0059	.0029
40	.0049	.0044	.0048	.0051	.0001
50	-.0019	.0046	-.0025	.0054	.0006

* $p<.0001$
** $p<.01$

40-item exam which had an absolute difference of .0004. In all cases, the standard errors of the means were slightly larger for the Rasch quick means than for the traditional means.

The second section of Table 3 contains the population means, standard errors of the means, and absolute mean differences for datasets 31 to 35. These datasets included the simulated responses of 100 persons to tests that had uniformly distributed item difficulties. Once again, there was a significant difference ($p<.0001$) between the traditionally calculated means and Rasch quick means for the 10-item test. The resulting values of the 2 means were very close for the 20- and 30-item tests with absolute differences of .0159 and .0032, respectively. There was no difference between the traditional and Rasch quick means for the 40-item test and hardly a discernable difference between the means for the 50-item test. Like datasets 1 to 30, the standard errors of the means for the traditional means were somewhat less than the Rasch quick means for all test lengths.

Datasets 36 through 40 contained simulated responses for 200 persons on tests with uniformly distributed item difficulties. Again, as seen in section 3 of Table 3, there was a significant difference ($p<.0001$) between the 2 types of means for the 10-item test. The smallest difference between the traditional and Rasch means was seen for the 40-item test which resulted in an absolute difference of .0002. The differences between the 2 differently-calculated means for the 30- and 50-item tests were very small with absolute differences of .0038 and .0010, respectively. The

difference between the traditional and Rasch quick means for the 20-item test, though larger than the differences seen for the 30-, 40-, and 50-item tests, was small with a value of .0126. As before, the Rasch standard errors of the means were slightly larger than the traditional means.

Section 4 of Table 3 consists of the population means, standard errors of the means, and absolute mean differences for datasets 41 to 45. These datasets included test responses for 300 persons who took exams with uniformly distributed item difficulties. Like datasets 1 to 40, a significant difference ($p<.0001$) was seen for the 10-item exam. The 2 kinds of means were very close for the other test lengths with the absolute differences ranging between .0107 and .0014 and decreasing as test length increased. As before, the Rasch standard errors of the means were larger than the traditional standard errors of the means for all test lengths.

Datasets 46 through 50 contained simulated responses for 400 persons who took tests with uniformly distributed item difficulties. The traditional mean and Rasch quick mean were significantly different ($p<.0001$) for the 10-item test with an absolute difference of .0732. Once again, the traditional and Rasch quick means for the 20-, 30-, 40-, and 50-item exams were very close with absolute differences ranging from .0128 down to .0001. The differences between the 2 types of means for the 40- and 50-item exams were hardly discernible; however, the 40-item exam yielded the smallest absolute difference of .0001 while the

50-item test resulted in an absolute difference of .0006. The standard errors of the means for the traditionally calculated means were smaller than those of the Rasch quick means.

Discussion

To determine if test length affected the difference between the traditionally calculated means and the Rasch quick means, the t-test results from the 10-, 20-, 30-, 40-, and 50-item tests were compared for each of the 5 sample sizes and 2 distributions of item difficulties. First, the 10-, 20-, 30-, 40-, and 50-item tests with normally distributed item difficulties were compared for the groups of 50 examinees. Next, the 5 test lengths with normally distributed item difficulties were compared for the samples of 100 persons and so on for each sample size and distribution.

It should be noted that there was a significant difference between the traditional and Rasch quick means for all of the 10-item tests with μ differences that ranged from .08 to .1. As for the 20-item tests, the absolute differences ranged between .1 and .2, which were noticeably less than the differences between the means for the 10-item tests. Except for one case, dataset 3, all of the absolute differences were less than .01 for the 30-, 40-, and 50-item tests. Thus, the differences were hardly discernible for these test lengths.

To ascertain if sample size affected the difference between the traditionally calculated means and the Rasch quick means, the results of the independent t-tests for experiments containing

sample sizes of 50, 100, 200, 300, and 400 were compared for each of the 5 test lengths and for each distribution. For example, first, samples sizes of 50, 100, 200, 300, and 400 were compared for the 10-item tests with normally distributed item difficulties. Next, the 5 sample sizes were compared for the 10-item tests with uniformly distributed item difficulties. This procedure was continued for each test length and distribution.

No trends were noted related to sample size. In general, the absolute differences were very similar across test lengths for each sample size.

To ascertain if the distribution of item difficulties affected the difference between the 2 types of means, the experiments with normal and uniform item difficulty distributions were compared with each of the 5 sample sizes and the 5 test lengths. First, the 2 distributions were compared for a sample of 50 persons and a 10-item test. Next, the distributions were compared for a sample of 100 persons and a 10-item test. This procedure was followed for each sample size and test length. With one exception, the absolute μ differences for the 20-item tests consisting of uniformly distributed item difficulties were less than those containing normally distributed item difficulties. As for the 30-item tests, the absolute μ differences were noticeably less for the tests that had uniform item difficulty distributions. No patterns related to type of item difficulty distribution were noted for the 10-, 40-, or 50-item tests.

Conclusions

The results of the study indicate that the Rasch quick norm procedure is a viable alternative to traditionally calculated norms. More specifically,

1. The Rasch quick norm procedure yielded means that were equivalent to traditionally calculated means for tests with a minimum of 30 items given to groups of 50 or more examinees. This held true for both normally and uniformly distributed item difficulties.

2. The 2 methods for calculating means were not equivalent for tests containing 10 items.

3. Although the absolute differences between the 2 types of means were less pronounced for the 20-item tests than for the 10-item tests, the differences were more pronounced than those of the 30-, 40-, and 50-item tests.

4. Sample size was not a determining factor in the differences between the traditional and Rasch quick means.

5. The absolute μ differences for datasets with uniformly distributed item difficulties were less than those with normally distributed item difficulties for the 20- and 30-item tests. No other patterns based upon item difficulty distribution were noted.

6. In all cases, the Rasch standard errors of the means were slightly larger than the traditional standard errors of the means.

Further Study

The quick norm procedure should be further investigated to determine its feasibility in situations other than those included in this study.

1. The norming method should be tested on populations that are not normally distributed.

2. The technique should be tested with exams which vary in overall difficulty. For example, the quick norm procedure could be studied for tests that are very easy, very hard, or of moderate difficulty.

3. It would be interesting to note how the quick norms would be affected by tests containing bias, guessing, or items with varying discrimination values.

Recommendations

The Rasch quick norm procedure is strongly recommended in cases where there is an existing bank of Rasch calibrated items. The simplicity and ease of use of the Rasch procedure is a decided advantage. The test user needs only 2 numbers: the frequency of persons who answered each item correctly and the Rasch-calibrated item difficulty, usually a part of an existing item bank. Norms can be computed quickly for any specific group of interest. In addition, once the selected items from the calibrated bank are normed, any test, built from the item bank, is automatically norm-referenced. Thus, the results of this study show that the quick norm procedure is a meaningful alternative to traditional "true score" norming for test users who desire normative data.

REFERENCES

Allen, M.J. & Yen, W.M. (1979). Introduction to measurement theory. Belmont, CA: Wadsworth.

Andrich, D (1988). Rasch models for measurement. Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-068. Beverly Hills and London: Sage Publications.

Crocker, L. & Algina, J. Introduction to classical and modern test theory. New York: Holt, Rinehart, and Winston.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

Linacre, J. M. (1992). BIGSTEPS. A Rasch test calibration computer program.

Luppescu, S. (1992). SIMTEST 2.1. A computer program for simulating test data.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), Educational Measurement, (3rd ed., pp. 221-262). New York: Macmillan.

Rasch, G. (1966). An item analysis which takes individual differences into account. The British Journal of Mathematical and Statistical Psychology, 19, 49-57.

SPSS-X User's Guide (3rd ed.). (1988). Chicago: SPSS Inc.

Wright, B. D. (1967, October). Sample-free test calibration and person measurement. Paper presented at The Invitational Conference on Testing Problems.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-115.

Wright, B. D. & Stone, M. H. (1979). Best Test Design. Chicago: Mesa Press.